



HAL
open science

A meta-analysis of infants' word-form recognition

Maria Julia Carbajal, Sharon Peperkamp, Sho Tsuji

► **To cite this version:**

Maria Julia Carbajal, Sharon Peperkamp, Sho Tsuji. A meta-analysis of infants' word-form recognition. *Infancy*, 2021, 26 (3), pp.369-387. 10.1111/inf.12391 . hal-03445953

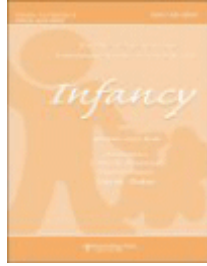
HAL Id: hal-03445953

<https://hal.science/hal-03445953>

Submitted on 6 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Maria Julia Carbajal, Sharon Peperkamp & Sho Tsuji

A meta-analysis of infants' word-form recognition

Journal:	<i>Infancy</i>
Manuscript ID	HIFC-2020-0069.R1
Manuscript Type:	Research Article
Keywords:	head-turn preference, meta-analysis, lexical acquisition, language acquisition, word-form recognition

SCHOLARONE™
Manuscripts

Abstract

Recognizing word-forms is an important step on infants' way towards mastering their native language. The present study takes a meta-analytic approach to assess overarching questions on the literature of early word-form recognition. Specifically, we investigated the extent to which there is cross-linguistic evidence for an early recognition lexicon, and how it may be influenced by infant age, language background, and familiarity of the selected stimuli (approximated by parent-reported word knowledge). Our meta-analysis - with open data access on metalab.stanford.edu - was based on 32 experiments in 16 different published or unpublished studies on infants 5-15 months of age. We found an overall significant effect of word-form familiarity on infants' responses. This effect increased with age, and was higher for infants learning Romance languages than other languages. We further found that younger, but not older, infants showed higher effect sizes for more familiar word lists. These insights should help researchers plan future studies on word-form recognition.

A meta-analysis of infants' word-form recognition

At the beginning of their second year of life, infants of all language backgrounds start producing words. This milestone in language development implies that by their first birthday, infants have come to understand at least some words. In order to achieve this, they must, firstly, segment word-forms out of continuous speech, secondly, store these word-forms in a recognition lexicon, and, thirdly, pair them to a meaning. A lot of research has focused on the onset of word learning in the first year of life (Jusczyk, 2002). Here, we are interested in the second step, word-form recognition. A seminal paper by Hallé & de Boysson-Bardies (1994) provided evidence for the emergence of a recognition lexicon in French-learning infants before the end of their first year of life. Infants were exposed in the headturn preference procedure (HPP; Kemler Nelson et al., 1995) to two types of word lists, one containing presumably familiar words, such as *ballon* ('ball'), *lapin* ('bunny'), and *chaussure* ('shoe'), and one containing presumably unfamiliar words, such as *caduc* ('obsolete'), *license* ('license'), and *volute* ('volute'). A familiarization phase was followed by a test phase, in which both 11- and 12-month-olds were found to listen longer to the familiar words. (The effect was larger for 12- than for 11-month-olds). In a second experiment, 11-month-olds were tested with stimuli that were better controlled for phonotactic complexity; longer listening times for familiar words were again observed.

Over the years, Hallé & de Boysson-Bardies's methodology has been used with infants learning a variety of languages. In the large majority of these studies, the aim went beyond the question of the age at which a recognition lexicon emerges. For instance, some investigated to what extent infants can recognize familiar words spoken in an unfamiliar accent (Best et al., 2009; Van Heugten & Johnson, 2014). Several other studies focused on

1
2
3 the specificity of early word-form representations, by examining which types of
4
5
6 mispronunciation interfere with familiar word recognition (Hallé & de Boysson-Bardies,
7
8 1996; Vihman et al., 2004; Swingley, 2005; Poltrock & Nazzi, 2015; Vihman & Majorano,
9
10 2017). Most studies, though, report at least one experiment showing the basic effect, i.e. a
11
12 familiarity preference for familiar words. In addition to French-learning infants, this basic
13
14 effect has thus been reported for 11-month-old infants learning English (Vihman et al.,
15
16 2004), Dutch (Swingley 2005), or Italian (Vihman & Majorano, 2017). There is one exception:
17
18 Vihman, Thierry, Lum, Keren-Portnoy, and Martin (2007) tested four groups of Welsh-
19
20 learning infants between the ages of 9 and 12 months and found no effect in any of the age
21
22 groups (although in the same study 11-month-old English-Welsh bilingual infants did show
23
24 the effect in both their languages). Results from our own lab also reveal mixed findings, with
25
26 some -unpublished- experiments (ANONYMIZED FOR REVIEW 1, Experiment 1 in
27
28 ANONYMIZED FOR REVIEW 2) replicating the effects found in the seminal study by Hallé &
29
30 de Boysson-Bardies' (1994), and others failing to do so (Experiment 2, ANONYMIZED FOR
31
32 REVIEW 2).

33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
These inconsistent findings led us to inquire further into the assumed mechanisms
underlying infants' developing preference for familiar word-forms. In our own lab, we had
previously started to investigate this question by assessing to what extent the effect is
based on a preference for speech sound sequences infants are frequently exposed to.
According to this hypothesis, infants would be expected to extract highly frequent sound
sequences from continuous speech and store them in a receptive protolexicon, which
should hence contain both real words (e.g. *ballon* 'ball') and other strings (e.g. *c'est pour*
'it's for'). Indeed, our experiments provided evidence for this hypothesis; most notably, no
listening time difference between high-frequency words and high-frequency nonwords

(ANONYMIZED FOR REVIEW 3). Consistent with other literature, these results demonstrate that infants are sensitive to the frequency structure of their input when building their language representations, be it on the phoneme (e.g., Tsuji & Cristia, 2014), word (e.g., Weisleder & Fernald, 2013), or sentence level (e.g., Mintz, 2003). The underlying assumption is that infants accumulate evidence about language while being exposed to environmental speech input. Acquisition on one level of representation has been suggested to influence others; for instance, word knowledge might help phoneme acquisition (Feldman, Griffiths, Goldwater, & Morgan, 2013; Swingley, 2009). Therefore, it is important to understand how exactly infants accumulate such evidence on a particular level of representation.

On the level of word-form recognition, experiments so far have rarely provided insight into how such evidence accumulation takes place. Would it make a difference whether an infant had encountered a speech string ten or one hundred times in order to obtain a preference over a never encountered one? The nature of infant experiments, in which it is hard to contrast more than two conditions (e.g., high-frequency versus low-frequency items like in ANONYMIZED FOR REVIEW 3; or native versus non-native items, see Tsuji & Cristia, 2014), makes answering this question particularly difficult. As noted by Swingley (2005), one additional drawback of the particular paradigm used for assessing familiarity preferences, which uses lists of items in each condition, is that we do not know how many of the familiar words infants should recognize in order for them to show a listening preference, nor can we infer which words exactly they recognize, regardless of whether they show the expected familiarity preference or not. Gaining insights into the way in which the items on these lists affect infants' familiarity preference can help us distinguish different assumed mechanisms. For instance, if the familiarity preference gets stronger in

1
2
3 relation to a stronger median word frequency, this might speak to gradual evidence
4
5 accumulation. If, however, some highly frequent items were driving the effect, this might
6
7 suggest a more threshold-like mechanism, in which only items for which infants have
8
9 encoded the meaning affect their familiarity preference.
10
11

12
13 Meta-analysis is an ideal tool to start answering questions that surpass what can be
14
15 measured in a single infant experiment. To provide some examples, by cumulating evidence
16
17 across a number of studies on early vowel discrimination, a recent meta-analysis has
18
19 demonstrated that infants' discrimination of two vowels gets stronger the further they are
20
21 acoustically apart (Tsuji & Cristia, 2017). Similarly, the larger the phonological difference
22
23 between a correct target noun and a mispronounced version of that noun, the better
24
25 infants discriminate the two (Von Holzen & Bergmann, 2018). In both cases, individual
26
27 studies did not necessarily compare infants' response to stimuli pairs with different acoustic
28
29 or phonological distances. Nevertheless, by coding these distances for each study and
30
31 comparing the strength of infants' responses across studies, meta-analytic techniques
32
33 enabled the authors to assess these questions about graded differences in infant responses.
34
35
36
37
38
39

40 Coming back to the topic of familiar word preferences, the mixed findings from our
41
42 own lab led us to question the role of the degree of familiarity, which might differ
43
44 depending on the frequency of the specific stimuli used in a particular study. To quickly sum
45
46 up our own stimulus selection process, our successful replication study of Hallé & de
47
48 Boysson-Bardies' original study involved novel recordings of the items used in their
49
50 Experiment 2 (ANONYMIZED FOR REVIEW 1). For the second study in which we used their
51
52 methodology, we replaced about half of both the familiar and unfamiliar items, and hence
53
54
55
56
57
58
59
60

again made new recordings.¹ This modification still led to a replication of the original effect with an effect size of $d = 0.63$ (ANONYMIZED FOR REVIEW 2). These stimuli were intended to be used next with bilingual infants; however, as we realized that several items were cognates in one or more languages that some of the bilingual infants we would recruit were likely be exposed to, we once more made a few replacements and re-recorded a final set. This time, and despite well-matched syllable structures and diphone frequencies across the lists of familiar and unfamiliar items, we failed to observe a listening preference altogether ($d = 0.05$; ANONYMIZED FOR REVIEW 2).

In an attempt to understand these mixed findings, we assessed the characteristics of our stimuli more closely. In choosing the items for our own experiments, we used frequency in child-directed speech from CHILDES corpora (MacWhinney, 2000) as a proxy for word familiarity². According to this measure, there was no difference in familiarity between the two sets of familiar words (i.e., the first and second experiment in ANONYMIZED FOR REVIEW 2, 2017). Thus, this rough familiarity estimate could not explain why we obtained a listening preference in one but not the other experiment. We noted, however, that the mean frequency of familiar items in both sets was almost 50% lower than that of the familiar items used in the second experiment of Hallé & de Boysson-Bardies (1994) and its replication with new recordings in ANONYMIZED FOR REVIEW 1. We also noted that the frequency of one of the items in the latter, *encore* 'more, again', was extremely high, i.e.,

¹ The replaced familiar items were *encore* 'again, more' and *bonjour* 'hello' because they often occur in isolation, *lapin* 'rabbit' because it contains the embedded word *pain* 'bread', *oiseau* 'bird' because it begins with a semi-vowel, and *biberon* 'bottle' because it contains an internal onset cluster. Lists of unfamiliar items were matched to the new familiar lists in terms of syllable structure and diphone frequencies.

² Details on our analyses can be accessed in our online material on the OSF in the document SI_1_CorpusAnalysis.docx

1
2
3 higher than 3 times the standard deviation above the mean; without this item, the mean
4
5 frequency was in fact not different from that in either one of our own sets.
6
7

8 Overall, two questions emerged. First, to what extent is there cross-linguistic
9
10 evidence for an early recognition lexicon, and is this effect robust beyond the age of 11
11
12 months? We were specifically interested in including unpublished data from other
13
14 researchers, which might shed more light on the reliability of the familiarity preference
15
16 effect in word recognition experiments. Second, how does word familiarity impact infants'
17
18 performance in this type of experiment? Since word familiarity is hard to assess, we used
19
20 parent-reported word knowledge as a proxy. The present meta-analysis is meant to answer
21
22 these and other questions. It is based on 32 experiments in 16 different studies, including 9
23
24 peer-reviewed articles. The experiments report on monolingual infants learning 7 different
25
26 languages. We will compute the effect size and analyze if there is evidence for a publication
27
28 bias, examine a possible correlation with *how* familiar the words in the familiar list are,
29
30 consider cross-linguistic differences, and investigate whether the effect changes with the
31
32 infant's age.
33
34
35
36
37
38
39

40 Methods

41 Preregistration and Open Access

42
43 We preregistered the analysis reported below on the Open Science Framework (OSF,
44
45 project link https://osf.io/6ty7b/?view_only=1228b7c023554b30b779f5a878b84579). This
46
47 preregistration was performed after collection of the meta-analytic dataset, but before data
48
49 inspection and analyses. The OSF project also contains documentation on the systematic
50
51 literature search process, search results and inclusion decisions, as well as analysis scripts
52
53 and supplementary information (henceforth referred to as SI).
54
55
56
57
58
59
60

Our analyses and visualizations are partly based on the scripts available on MetaLab (<https://metalab.stanford.edu>), an online repository for meta-analyses on infant language development (Bergmann et al., 2018). To run these analyses, we used the packages metafor version 2.1-0 (Viechtbauer, 2010) and the tidyverse version 1.2.1 (Whickham, 2017) in R version 3.5.3 (R Core Team, 2019) and R Studio version 1.1.456 (R Studio Team, 2019).

A static version of the meta-analytic dataset corresponding to the dataset as of submission of the present manuscript is available on our OSF project. In addition, MetaLab contains a dynamic dataset, which will be updated each time new data consistent with the inclusion criteria of the present meta-analysis become available. Future readers of this manuscript are thus encouraged to apply our analysis scripts to any updated future dataset in order to obtain the most up-to-date meta-analytic estimates.

Systematic Literature Search

The search followed the PRISMA protocol (Moher, Liberati, Tetzlaff, Altman, & the PRISMA Group, 2009), and included search, abstract and fulltext screening conducted by the first author of the present study.

Studies included in our meta-analysis had to fulfill the following criteria:

- (1) They were conceptual replications of the seminal paper by Hallé and de Boysson-Bardies (1994).
- (2) They had as experimental design within-subject comparisons of attention to lists of familiar words versus lists of novel or rare words.
- (3) They studied monolingual infants between 0-15 months of age.
- (4) They presented word lists in the infants' native language.
- (5) They used behavioral or electrophysiological measures to study word recognition.
- (6) They could be either peer-reviewed or not.

(7) They did not include novel words that were phonological neighbors of familiar words.

(8) They used t-tests between raw looking time data of two conditions, or we were able to retrieve t-tests results with the help from the authors.

(9) If they included an exposure phase, this phase immediately preceded the test phase.³

We conducted our literature search between October 2014 and December 2017.

Since the meta-analysis was originally motivated by the first author's attempt to replicate previous studies, a first list of candidate papers was already present at the beginning of the search phase in October 2014. This list had been assembled based on seminal articles and references therein, suggestions from experts, and a brief database search. Additional candidates were then detected during a conference attended by the first author in 2017. A systematic database search was conducted in December 2017 using Google Scholar with the keywords "infant word-form recognition". This search was supplemented by a seed search of papers citing the first published article on the topic (Hallé & de Boysson-Bardies, 1994), and by checking the references in review papers found during abstract screening of the main database search. Finally, one additional candidate study was identified while the second author was serving on a thesis committee.

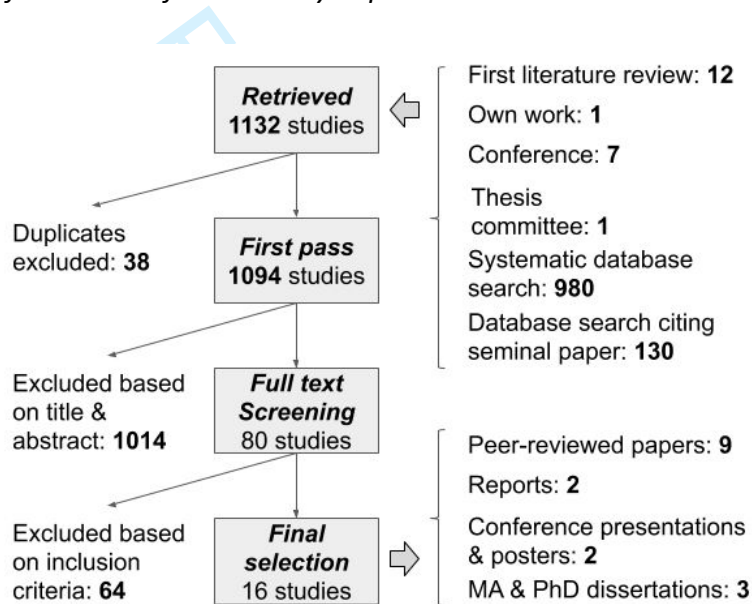
After our search, we contacted the authors of 8 studies, because their article or dissertation were not accessible (2 studies), or data necessary to conduct the meta-analysis were missing (8 studies). All authors answered, and we succeeded in retrieving the missing information in all but one case. The search results are summarized in a PRISMA flowchart

³ Thus, we did not include, for instance, Jusczyk and Hohne (1997), where infants received home exposure for several days before being tested in the lab.

(Fig.1), and more extensively documented in a spreadsheet accessible in the SI. Our final search sample included 16 studies.

Figure 1

PRISMA flow-chart of meta-analytic process.



Sample Description

Of the 16 studies in the final sample, 9 were peer-reviewed, 2 were reports, 2 were conference presentations or posters, and 3 were MA or PhD dissertations. The studies were reported between 1994 and 2019 by 13 different first authors. These studies included a total of 32 experiments that provided unique effect sizes for our meta-analysis.

The methodology varied in several ways across the included experiments, but always contained the main feature of that in Hallé & de Boysson-Bardies (1994), i.e. the presentation of alternating lists of familiar and unfamiliar items. For instance, word type per condition varies between 10 and 33 types, the range for tokens per trial is 11-24, and the number of trials per condition varies from 3 to 8. Some omitted the familiarization phase

1
2
3 (and among these, some excluded the first few test trials from analysis). Finally, some
4
5 included highly familiar words like *mummy* and *hello* (or the equivalent in the relevant test
6
7 language), while others did not.
8
9

10 Of the included experiments, nine were conducted with infants learning British
11
12 English, eight with infants learning French, six with infants learning Welsh, three with infants
13
14 learning Spanish, two with infants learning Japanese, two with infants learning North
15
16 American English, one with infants learning Dutch, and one with infants learning Italian. All
17
18 experiments were behavioral, with 25 of them using the head-turn preference procedure
19
20 (HPP, see Kemler Nelson et al., 1995), and 7 using central fixation (CF, see Werker, Cohen,
21
22 Lloyd, Casasola, & Stager, 1998). Some had additional EEG recordings, which are not used in
23
24 the present analysis. Twenty-two of the experiments included a short familiarization phase,
25
26 while the remaining ten only had a test phase.
27
28
29
30
31

32 The independent variable in all of the studies was whether the word lists presented
33
34 were composed of familiar words or novel or rare words, and the dependent variable was
35
36 infants' mean looking time to the respective lists. The experiments served as a baseline
37
38 experiment for other experiments in 12 cases, and the main experiment in its own right in
39
40 18 cases; for the two remaining ones this was not specified.
41
42
43
44

45 **Effect Size Computation**

46
47 We calculated Hedges' *g*, a variant of the standard Cohen's *d* effect size that corrects
48
49 for small sample sizes (Hedges, 1981). In 30 cases, we were able to calculate it based on the
50
51 means and standard deviations provided either in the manuscript, in the figures, or by the
52
53 authors. In the remaining two cases, we calculated the effect size based on reported *t*-
54
55 values. We based our effect size calculations on the scripts provided in MetaLab (Bergmann
56
57 et al., 2018). We coded all effect sizes such that a positive value would indicate a familiarity
58
59
60

1
2
3 preference, thus a preference for the familiar words, and a negative value would indicate
4
5 the opposite preference for the novel or unknown words.
6
7

8 In order to calculate the standard error of effect sizes based on within-subject
9
10 comparisons, it is necessary to know the correlation between the two measurement points.
11
12 We were able to obtain these correlations from the authors or from the reported t-values in
13
14 29 cases. In two further cases, we converted the reported F value to a t-value using the
15
16 formula $t = \sqrt{F}$, which was justified since the F-test only compared the two groups of
17
18 interest. Finally, for one datapoint, we imputed the missing correlation by sampling
19
20 randomly from a normal distribution with the median and variance of the known
21
22 correlations.
23
24
25
26

27 **Coding of Moderator Variables**

28
29 We coded three main moderator variables for inclusion in our meta-analytic model.
30
31 First, we coded infant native language, information that was obtained from the articles or
32
33 authors. We subsequently created the new variable language group in order to reduce the
34
35 number of levels of this variable for the purpose of analysis. We grouped together French,
36
37 Italian, and Spanish into Romance languages; British English, Canadian English, North
38
39 American English, and Dutch into Germanic languages; Welsh and Japanese were not
40
41 further grouped.
42
43
44
45
46

47 We secondly coded infant age in days, which was either directly reported, obtained
48
49 from authors, or converted from age in months by multiplying with 30.42.
50
51

52 Obtaining a proxy for our third moderator variable, word familiarity, required a few
53
54 more steps. We first assembled a list of the words used in the familiar word lists in each
55
56 study. In order to approximate how familiar they would be for an infant with a given native
57
58 language, we then looked up infants' knowledge of these words according to parental
59
60

1
2
3 report in each language's equivalent of the MacArthur Bates Communicative Development
4
5 Inventory (CDI; Fenson, 2007). Note that this measure is quite different from the CHILDES
6
7 counts used to evaluate item familiarity in our own experimental studies, as mentioned in
8
9 the Introduction: the former are parental reports based on a finite number of possible word
10
11 candidates, while the latter are derived from word counts in natural conversations. Here,
12
13 we based ourselves on the CDI counts for various reasons. First, CDI counts were more
14
15 consistently available for all languages for which we needed to obtain these counts. Second,
16
17 even if CHILDES data were available, these corpora vary along many dimensions across
18
19 languages, and thus might not be suitable for the direct comparisons we attempted to make
20
21 in the present study. One might object that although CHILDES data might be considered
22
23 more representative since they are based on transcriptions of natural conversations, a
24
25 drawback is that they are often based on a very small number of children, while CDI scores
26
27 are available across a larger number of participants. Note, however, that frequencies
28
29 derived from both CHILDES corpora and CDI reports are suboptimal in the sense that they
30
31 reflect grouped statistics that do not necessarily reflect infants' individual experience with
32
33 words.
34
35
36
37
38
39
40
41

42 To obtain CDI scores, we turned to WordBank (Frank, Braginsky, Yurovsky &
43
44 Marchman, 2016), an online repository of these vocabulary questionnaires. Since WordBank
45
46 relies on the submission of data sets by individual researchers, data for all applicable age-
47
48 groups in all tested languages are not uniformly available. We therefore instead aimed to
49
50 obtain questionnaire data from infants at or around 11 months of age, the age-group tested
51
52 in the majority of studies in our sample (14 experiments at 11 months, 12 other
53
54 experiments at 10 months or 12 months of age). Data availability on WordBank for the
55
56 languages we required was highest for 12-month-olds, and we therefore decided to
57
58
59
60

1
2
3 assemble data from that age group. We were able to obtain data from French, Italian,
4
5 Spanish, and English, thus covering 24 of the included experiments. We were further able to
6
7 obtain vocabulary data from 12-month-old Dutch infants via the Baby Research Center in
8
9 Nijmegen, whose data were anonymously shared with us (2 experiments). For Japanese
10
11 data (2 experiments), we relied on the norming data for 12-month-aged infants reported in
12
13 Ogura, Watamaki, and Inaba (2016). We were not able to obtain questionnaire data for
14
15 Welsh (6 experiments). For each of the familiar words in the experimental lists, we checked
16
17 whether it was included in the vocabulary questionnaires, and if yes, we coded the
18
19 percentage of infants that were reported to comprehend the word. In some cases, the word
20
21 in the word list differed slightly from the word in the questionnaire (for instance, British
22
23 English: *telly* in word list, *TV/Television* in questionnaire; Japanese: *haitta* ('it went in') in
24
25 word list; *hairu* ('it goes in') in questionnaire). We decided to include these words in our
26
27 counts. Finally, we computed the median and maximum percentages of comprehended
28
29 words over all items in the familiar word list for a given experiment. We computed these
30
31 two indices in order to account for two possible ways in which word familiarity could
32
33 influence infants' cumulative looking times over a trial. That is, infant looking times might
34
35 either be a function of the median familiarity of words in a list, or of one or more very high-
36
37 familiarity items. In case a word was not included in vocabulary questionnaires, it was
38
39 treated as a missing value. Overall, for the 23 word lists included in the present study, all
40
41 words were available for 14 lists, 95% for one, 92% for four, 72% for three, and 38% for one
42
43 list.

44
45 In addition to these main moderator variables, we also coded standard
46
47 methodological variables, including whether the experiment had an exposure phase prior to
48
49 test, and which testing method was used. We did not include those variables in the
50
51
52
53
54
55
56
57
58
59
60

analyses, mainly because studies were not evenly distributed across the categories of these variables. They are, however, accessible on the OSF project page and in the MetaLab dataset.

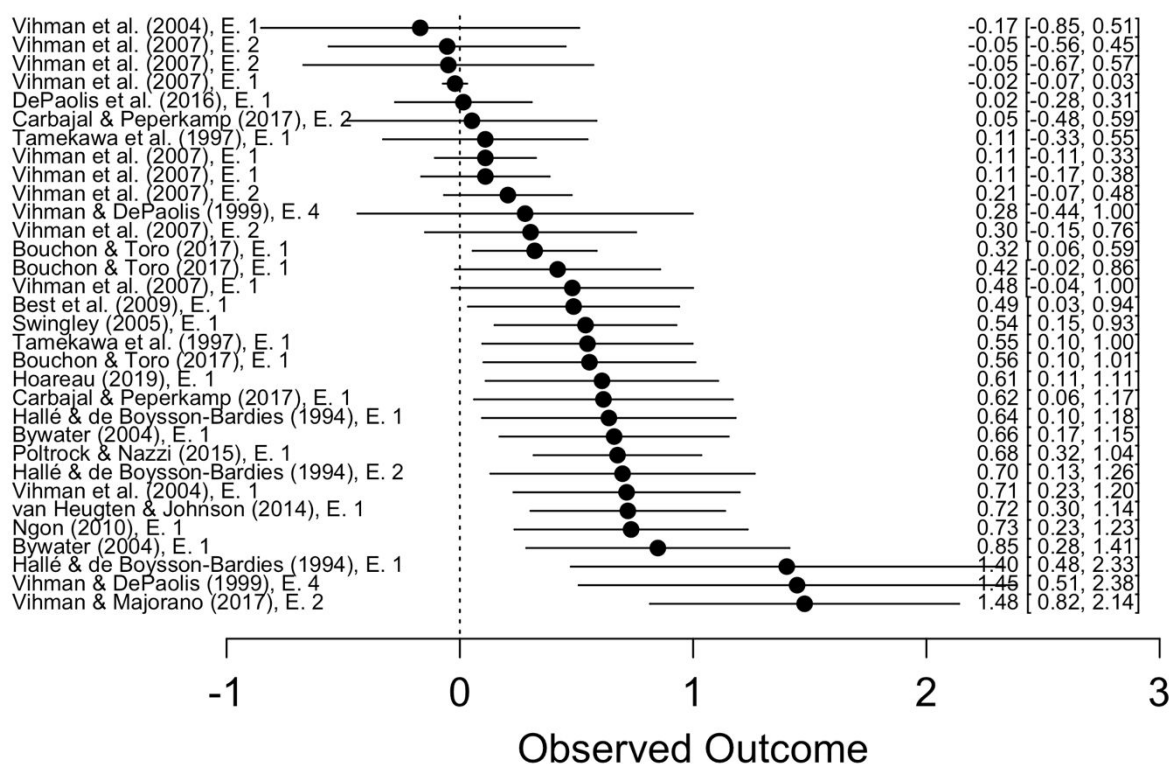
Before analysis, we centered the continuous predictors infant age as well as median and maximum word familiarity, and sum-coded the contrasts for the predictor variable language group.

Results

An overview of the experiments included and their associated effect sizes can be found in the forest plot in Figure 2, as well as in Table 1.

Figure 2

Forest plot of experiments and associated effect sizes, ordered by effect size magnitude.



1
2
3 *Note.* Dots represent effect size by experiment, and surrounding error bars indicate
4
5 standard errors of effect size. Mean and confidence intervals are provided on the right side
6
7 of the figure.
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review Only

INFANT WORD-FORM RECOGNITION META-ANALYSIS

Table 1

Overview of studies included in meta-analysis.

Citation	Peer reviewed	No. exp.	No. infants	Native language	Method	Familiariz. phase	Mean age (months)
Best et al. (2009)	Yes	1	20	English	CF	No	15
Bouchon & Toro (2017)	No	3	22/22/14	Spanish	CF	No	5/8/12
Bywater (2004)	No	2	24/12	English	HPP	Yes	10/11
ANONYMIZED 2	No	2	14/14	French	CF	No	11/11
DePaolis et al. (2016)	Yes	1	53	English	HPP	Yes	10
Hallé & de Boysson-Bardies (1994)	Yes	3	12/16/12	French	HPP	Yes	11/11/12
Hoareau (2019)	No	1	32	French	HPP	No	12
ANONYMIZED 1	No	1	16	French	CF	No	11
Poltrock & Nazzi (2015)	Yes	1	24	French	HPP	No	11
Swingley (2005)	Yes	1	24	Dutch	HPP	Yes	11
Tamekawa et al. (1997)	No	2	24/24	Japanese	HPP	Yes	10/12
van Heugten & Johnson (2014)	Yes	1	16	English	HPP	No	15
Vihman & DePaolis (1999)	No	2	12/12	Welsh	HPP	Yes	11/12
Vihman & Majorano (2017)	Yes	1	20	Italian	HPP	Yes	11
Vihman et al. (2004)	Yes	2	12/12	English	HPP	Yes	9/11
Vihman et al. (2007).1	Yes	4	25/27/23/26	English	HPP	Yes	9/10/11/12
Vihman et al. (2007).2	Yes	4	14/12/27/21	Welsh	HPP	Yes	9/10/11/12

Note. No. exp.: Number of experiments of this study included in meta-analysis. No. infants: Number of infants included in each experiment. CF

= Central Fixation, HPP = Head-turn preference procedure.

Effects of infant age and language background

All reported analyses follow our preregistered protocol. We first report the results for an intercept-only baseline random-effects model without moderators. For the random effects structure of this model, we nest each unique effect size under the study it comes from to account for the fact that effect sizes derived from the same studies might be more similar than those from different studies. The model took the form $rma.mv(g, se_g, random = \sim 1 | study/unique_es)$. The model estimate was $g = 0.5$, and differed significantly from zero [$se = 0.08, z = 6.39, p < .001, CI_l = 0.349, CI_u = 0.657$], thus demonstrating the presence of an effect of word familiarity.⁴ The Q-test for heterogeneity was significant [$Q(31) = 135.91, p < .001$], indicating that a significant portion of variance remained unexplained by this model.

We therefore moved on to the moderator analysis, where we now added language group (Romance, Germanic, Welsh, Japanese) and age as predictors. As to age, we constructed three age models, in which we respectively modeled linear, quadratic, and cubic effects of age, and performed model comparisons with likelihood ratio tests between these three models to determine the best model fit. The intuition behind not taking a linear increase of familiar word preference for granted is that infants might, at one point in development, stop preferring the familiar over unfamiliar items, either because they start developing a novelty preference (Hunter & Ames, 1988) or because they start perceiving the previously unfamiliar words as familiar (see also Vihman et al., 2007). The model comparison revealed no significant differences between the three models (linear-quadratic: $X^2(3) = 0.038, p = .99$; linear-cubic: $X^2(6) = 1.59, p = 0.95$; quadratic-cubic: $X^2(3) = 1.56, p = 0.67$). Since the linear model provided the best model fit (AIC, BIC, AICc), we decided to keep

⁴ As preregistered, we conducted the same analysis for the subset of 11-month-olds, the age group tested most frequently with this paradigm. This analysis resulted in a higher effect size estimate [$g = 0.61, se = 0.08, z = 7.92, p < .001, CI_l = 0.457, CI_u = 0.758$].

1
2
3 this model. The Q-test for heterogeneity in this model remained significant [$Q(24) = 50.84$, p
4
5 = 0.001]. The Q-test for moderators further indicated that the moderators explained a
6
7 significant portion of variance [$Q(7) = 15.89$, $p = 0.026$]. The model intercept was again
8
9 significant [$g = 0.39$, $se = 0.08$, $z = 4.95$, $p < .001$, $CI_l = 0.236$, $CI_u = 0.542$], as was the main
10
11 effect of age [$g = 0.004$, $se = 0.002$, $z = 2.39$, $p = .017$, $CI_l = 0.0007$, $CI_u = 0.0071$]. The mean
12
13 effect size for studies conducted in Romance languages was significantly different from the
14
15 mean [$g = 0.26$, $se = 0.10$, $z = 2.46$, $p = .014$, $CI_l = 0.0526$, $CI_u = 0.4633$], which was not the
16
17 case for the other language groups. No interaction effect was significant. The table of full
18
19 results can be accessed in the SI.
20
21
22
23

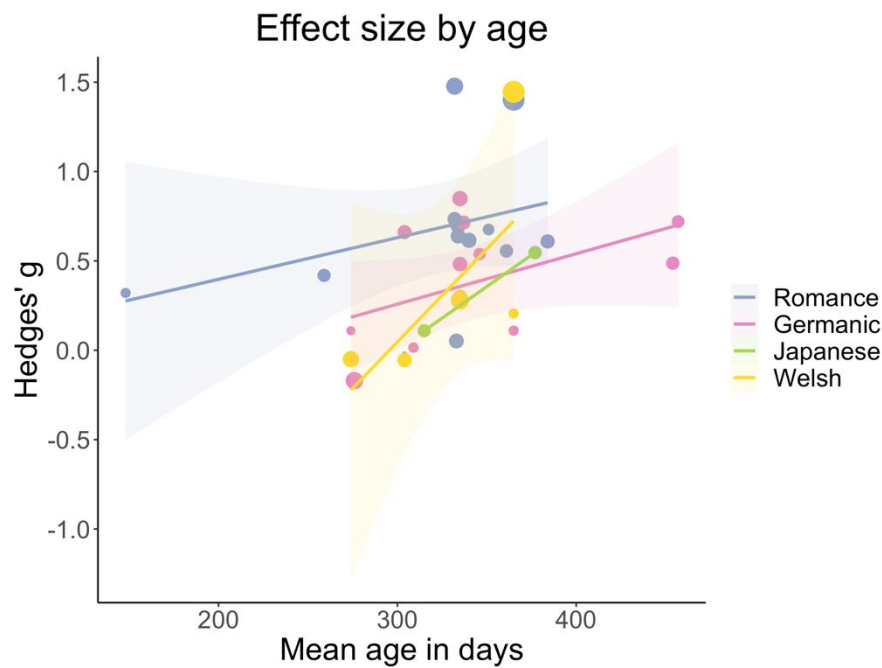
24
25 Since visual inspection of Figure 3 did not preclude the possibility of single
26
27 datapoints driving the age effect, we performed a leave-one-out analysis to make sure this
28
29 was not the case. In this analysis, we ran the regression model as many times as there were
30
31 datapoints (thus, 32 times) while removing one of the datapoints each time. If one of the
32
33 datapoints was driving the results, we should find no significant age effect in one or more of
34
35 the model runs, while a significant effect in all runs above the significance threshold of 95%
36
37 would suggest a robust age effect. An examination of p-values obtained by this procedure
38
39 shows no influence of single datapoints, with a range of p-values between $p = .005$ and p
40
41 = .028 (mean = .018). Results for the other variables also followed the pattern of the original
42
43 analysis. Full results are reported in the SI.
44
45
46
47
48

49 As preregistered, we additionally conducted the moderator analysis only for infants
50
51 at 11 months, the age group tested most frequently with this paradigm. For this purpose,
52
53 we took out the age predictor). This analysis continued to show a significant intercept [$g =$
54
55 0.53, $se = 0.09$, $z = 6.16$, $p < .001$, $CI_l = 0.362$, $CI_u = 0.699$], but no other significant effects.
56
57
58 This leaves it hard to interpret the effect of language group found before, especially so
59
60

because the distribution of effect sizes differed across age groups (see Fig. 3). The table of full results can be accessed in the SI.

Figure 3

Hedges' g effect size as a function of infant age and language family background.



Note. Point size indicates inverse effect size variance, with larger points being weighted stronger in the regression model. Lines and their shading represent linear fit and confidence intervals.

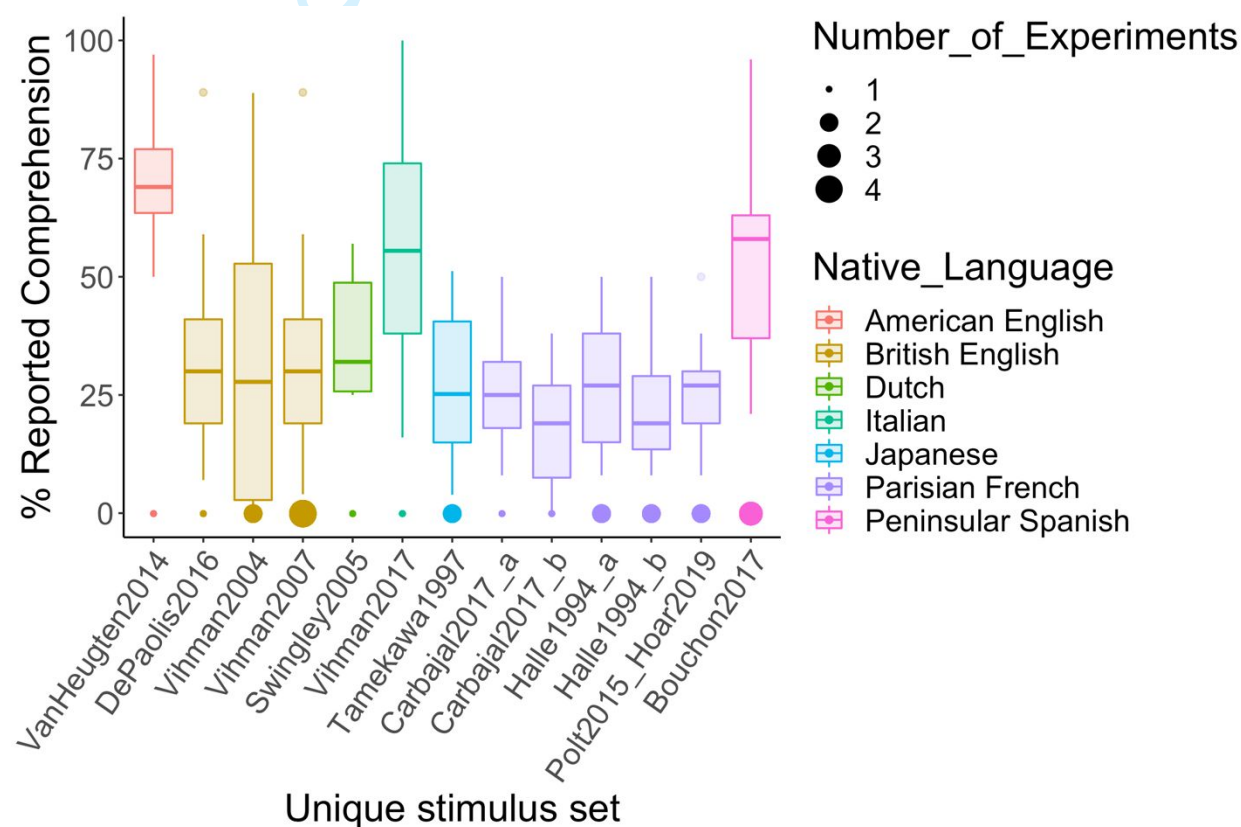
Effects of infant age and word familiarity

In our final analysis step, we assessed the influence of word familiarity on effect sizes. As described in the previous section, we took as a proxy for this predictor the median and maximum percentage of 12-month-old infants that were reported in large scale questionnaire data to know a given word from the familiar word lists. This computation revealed a large heterogeneity, with the median familiarity of the words in a given familiar

word list ranging between 19% and 69% (see also Fig. 4), and the maximum familiarity ranging between 38% and 100%.

Figure 4

Percentage of infants knowing a given word at age 12 months by word list.



In order to investigate how these differences in median or maximum familiarity would affect average infant looking times, we constructed two separate meta-analytic regression models. As in the previous moderator model, we added infant age and the interaction of age and familiarity as moderators. We did not add language group to this analysis in favor of preserving power given the relatively small size of our dataset. We first

report results for the model including median word familiarity (Fig.5). As in the previous model, both the Q-test for residual heterogeneity [$Q(19) = 50.84, p < .001$] and the Q-test for moderators [$Q(3) = 11.40, p = 0.010$] were significant. Further, the model intercept was significant [$g = 0.447, se = 0.066, z = 6.74, p < .001, CI_l = 0.317, CI_u = 0.577$], as was the main effect of age [$g = 0.005, se = 0.002, z = 3.012, p = .003, CI_l = 0.002, CI_u = 0.008$]. The main effect of median word familiarity approached significance [$g = 0.908, se = 0.479, z = 1.897, p = .058, CI_l = -0.030, CI_u = 1.845$], indicating larger effect sizes with higher word familiarity. Finally, the interaction between age and word familiarity was significant [$g = -0.014, se = 0.007, z = -2.07, p = .039, CI_l = -0.028, CI_u = 0.039$], indicating that the effects of word familiarity on effect sizes decreased with age.

One possible concern about this analysis is that the median values might be skewed: Although in most cases we were able to obtain familiarity scores for all or most words in a given list, this was not the case for all of them. While these cases were treated as missing values, and thus did not influence the median, it is conceivable that the actual familiarity values of these items would be quite low, given their lack of representation in standard vocabulary lists. Based on a reviewer's suggestion and in order to assess whether including those items with low values would change the results, we conducted an additional exploratory analysis. In this analysis, we replaced each of the missing values with a random number drawn from the interval between 0 and the minimum familiarity score of the list to which the given item belongs. This analysis showed very similar results: A significant model intercept [$g = 0.495, se = 0.074, z = 6.684, p < .001, CI_l = 0.350, CI_u = 0.640$], a significant main effect of age [$g = 0.004, se = 0.001, z = 3.606, p < .002, CI_l = 0.002, CI_u = 0.006$], and a significant interaction between age and word familiarity [$g = -0.026, se = 0.010, z = -2.482, p = .013, CI_l = -0.046, CI_u = 0.006$]. The effect of median word familiarity, which had been

1
2
3 approaching significance in the previous analysis, now reached significance [$g = 2.338$, $se =$
4
5
6 0.946 , $z = 2.473$, $p = .013$, $CI_l = 0.485$, $CI_u = 4.191$].

7
8 As to the model including maximum word familiarity, the Q-test for residual
9
10 heterogeneity was significant [$Q(19) = 69.41$, $p < .0001$], while the Q-test for moderators
11
12 was marginally significant [$Q(3) = 6.96$, $p = 0.073$]. In this analysis, the model intercept [$g =$
13
14 0.427 , $se = 0.083$, $z = 5.15$, $p < .001$, $CI_l = 0.264$, $CI_u = 0.590$] and the main effect of age [$g =$
15
16 0.005 , $se = 0.002$, $z = 2.27$, $p = .023$, $CI_l = 0.001$, $CI_u = 0.009$] were significant, but no effects
17
18 involving maximum word familiarity showed a statistically significant effect. Detailed results
19
20 are reported in the SI⁵.
21
22
23

24
25 It is possible that familiarity effects are not driven by only the most familiar item, but
26
27 by a combination of multiple high-familiar items in a given list. Based on reviewers'
28
29 suggestions, we explored two alternative ways to enter highly familiar items in the analysis.
30
31 First, we replaced the maximum familiarity score by the mean of all items that had a
32
33 familiarity score of at least 1 SD above the mean of the word list in question. Second, we
34
35 replaced the maximum familiarity score by the mean of all items that were in the highest
36
37 25% quantile of familiarity scores of a given list. In both analyses, the results were
38
39 comparable to those obtained in the analysis based on one maximally familiar item, with a
40
41 significant intercept [SD analysis: $g = 0.448$, $se = 0.075$, $z = 6.015$, $p < .001$, $CI_l = 0.302$, $CI_u =$
42
43 0.594 ; Quantile analysis: $g = 0.460$, $se = 0.075$, $z = 6.136$, $p < .001$, $CI_l = 0.313$, $CI_u = 0.607$],
44
45 no effect of the familiar items [SD analysis: $g = 0.143$, $se = 0.383$, $z = 0.374$, $p = .708$, $CI_l = -$
46
47 0.607 , $CI_u = 0.894$; Quantile analysis: $g = 0.247$, $se = 0.402$, $z = 0.613$, $p = .540$, $CI_l = -0.542$,
48
49
50
51
52
53
54

55
56
57 ⁵ As preregistered, we also performed the same two analyses with the subset of 11-month-
58
59 old infants, which did not lead to any significant effects of word familiarity. These analyses
60 can be found in the SI.

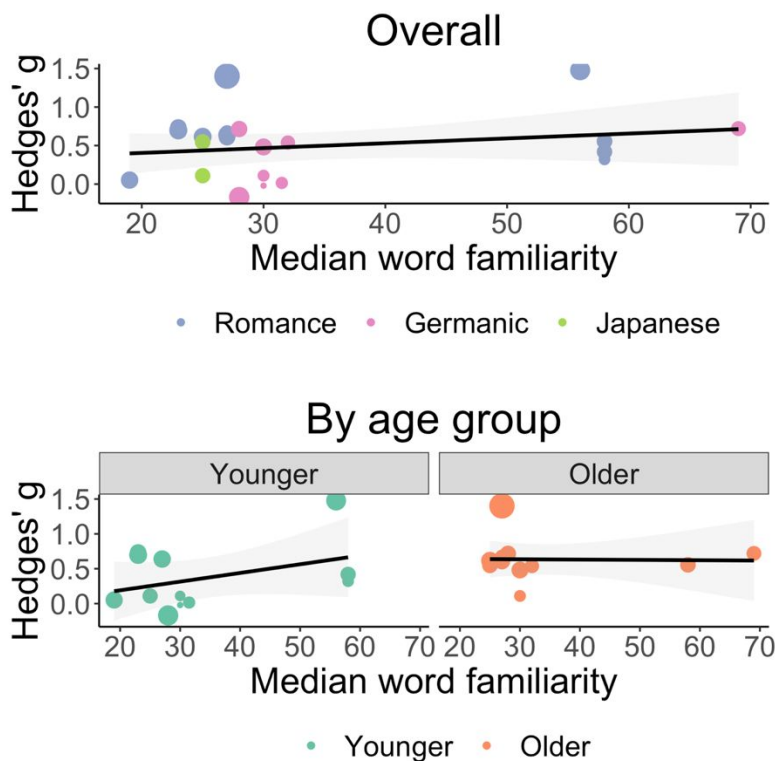
1
2
3 $CI_u = 1.035]$ and a significant effect of age [SD analysis: $g = 0.004$, $se = 0.001$, $z = 3.041$, p
4 $= .002$, $CI_l = 0.001$, $CI_u = 0.007$; Quantile analysis: $g = 0.004$, $se = 0.001$, $z = 2.838$, $p = .005$, CI_l
5 $= 0.001$, $CI_u = 0.006$]. The interaction between familiarity and age was marginally significant
6 in the SD analysis [$g = -0.011$, $se = 0.006$, $z = -1.785$, $p = .074$, $CI_l = -0.022$, $CI_u = 0.001$], but
7 not in the quantile analysis [$g = -0.009$, $se = 0.006$, $z = -1.498$, $p = .134$, $CI_l = -0.022$, $CI_u =$
8 0.029].

9
10
11
12
13
14
15
16
17
18 Finally, in order to again assess the possibility that a few studies were driving the
19 interaction effect, we conducted a leave-one-out analysis. In this case, the distribution of p-
20 values suggested that this could indeed be the case, with a range of p-values between p
21 $< .001$ and $p = .129$ (mean = $.047$). Of these, 22 of 32 p-values were below the threshold of p
22 $= .05$, 8 were below the threshold of $p = .10$, and 2 were above $p = .10$. The datapoints left
23 out when obtaining the two last, non-significant results were indeed at the lower and higher
24 end of the familiarity scores (28% and 69%, respectively), lending support to the hypothesis
25 that a few values were driving the effect, and cautioning us to avoid over-interpreting the
26 present findings.

27
28
29
30
31
32
33
34
35
36
37
38
39
40 Together, this last set of analyses shows **cautionary** preliminary evidence that
41 differences in familiarity of experimental items are indeed reflected in infant looking times,
42 such that infants look longer the higher the median familiarity of items in the familiar word
43 list is. Further, if anything, infants' looking times are driven by the median familiarity rather
44 than by the presence of one or more highly familiar items. Finally, the possible effect of
45 familiarity gets smaller with infant age. **More data are needed to confirm these patterns.**

46 47 48 49 50 51 52 53 54 **Figure 5**

55
56
57 *Effect sizes as a function of word familiarity and infant age.*
58
59
60



Note. Point size indicates inverse effect size variance, with larger points being weighted stronger in the regression model. In the bottom panel, infants were median-split by age for the purpose of visualization.

Publication Bias

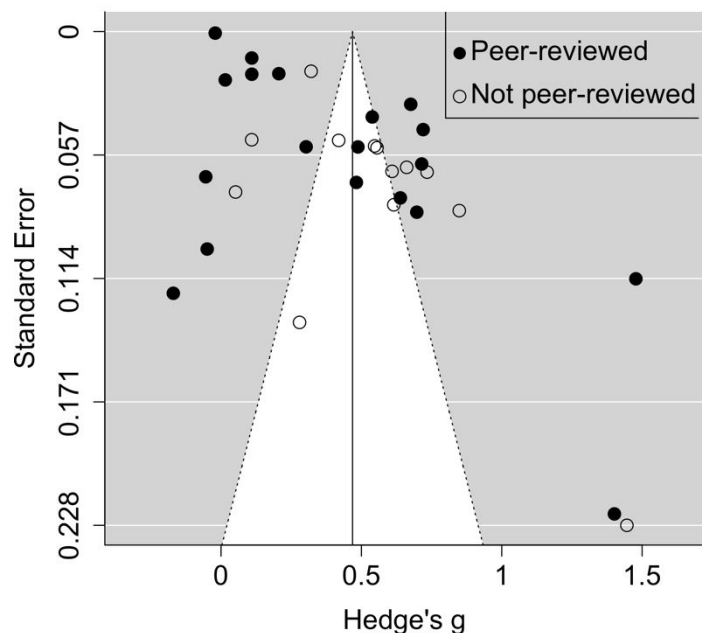
Meta-analysis can suffer from selective reporting. In particular, only those results that yielded effects in the expected direction might have been published. Our meta-analysis contains a relatively large proportion of data that are not published in a peer-reviewed journal, but this might not be sufficient to forego bias. A funnel plot of our data is shown in Figure 6. In funnel plots, effect sizes are plotted against their standard error as a measure of study size and study precision. Studies with higher precision are expected to be closer to the true effect size and thus cluster around the middle, while studies with lower precision are expected to spread to both sides. Asymmetric distribution of datapoints around the funnel are thus a potential indicator of publication bias. Egger's test for funnel plot asymmetry

1
2
3 revealed significant asymmetry ($z = 3.61, p < 0.001$), which was also true when looking at
4
5 only the peer-reviewed ($z = 2.52, p = 0.012$) or non-peer-reviewed ($z = 2.39, p = 0.012$)
6
7 records.
8
9

10 Visual inspection of the dataset suggested that the three datapoints in the lower
11
12 right corner of the plot might be driving the publication bias; and indeed, removing these
13
14 three datapoints eliminated this bias ($z = 0.61, p = 0.540$). We did, however, refrain from
15
16 removing the datapoints from further analyses, as this was not part of our preregistration. It
17
18 is worth noting that standard meta-analytic practice in MetaLab suggests removing
19
20 datapoints only if they are over 3 standard deviations away from the mean. All three
21
22 datapoints in questions were over 2, but below 3 standard deviations above mean. Thus,
23
24 even if we had preregistered the removal of outliers these datapoints would have remained
25
26 in the dataset. Finally, asymmetries (detected) in a funnel plot (by Egger's test) can occur for
27
28 a variety of reasons other than publication bias (Cochrane Collaboration, 2011). While we
29
30 [conducted](#) analyses with the full dataset, we invite the reader to keep the possibility of a
31
32 biased dataset in mind.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56

57 **Figure 6**

58
59 *Funnel plot of effect sizes against their standard error.*
60



Note. The vertical line in the middle indicates the estimated average effect size across studies, and the white area indicates a pseudo confidence interval with bounds equal to ± 1.96 standard errors.

Discussion

The present meta-analysis assessed the effect of word familiarity on infants' word-form recognition. Partly motivated by our own difficulties in consistently replicating this effect, our research question was twofold: First, does the literature overall yield evidence for early recognition of familiar word-forms? Second, what are the moderators that affect the strength of this effect?

Our base model confirmed a significant main effect of word-form familiarity: Across the 32 experiments entered into the meta-analysis, we found evidence for a medium effect size across all age-groups and language backgrounds assessed. That same analysis also showed that a significant portion of variance remained unexplained.

1
2
3 As preregistered, we therefore moved on to a moderator analysis, including age and
4 language group as predictors. As to age, we had foreseen several possible ways in which it
5
6 language group as predictors. As to age, we had foreseen several possible ways in which it
7
8 could affect word-form recognition: The effect could be linear, with better word-form
9
10 recognition with higher age. However, the literature has also suggested that the effect
11
12 might be non-linear, for instance due to an increase in preference for familiar word-forms
13
14 up to the age of around 11 months, and a switch to a novelty preference thereafter (see
15
16 Hunter & Ames, 1988; but Bergmann & Cristia, 2016). Our model showed a significant linear
17
18 increase of effect sizes with age, but no evidence that a model with a non-linear effect of
19
20 age would show a better fit. Our data thus suggest that infants' preference for familiar
21
22 word-forms increases with age. We do, however, note that the age distribution of our data
23
24 does not allow firm conclusions to that effect: Only 7 of 32 datapoints assess infants above
25
26 the age of 12 months.
27
28
29
30
31

32 Overall, infants from Romance language backgrounds (subsuming datapoints from 8
33 French, 3 Spanish, and 1 Italian native language backgrounds) showed higher effect sizes
34
35 than infants from other language backgrounds. We are cautious to interpret this finding,
36
37 since the distribution of datapoints across language backgrounds and infant age is far from
38
39 even. We did, however, follow up on one possible explanation for this effect, namely that
40
41 the word lists presented to infants from these language backgrounds contained more
42
43 familiar words than those presented to infants from other language backgrounds. To this
44
45 end, we used the metrics of percentage reported comprehension. We did not find evidence
46
47 for this possibility, since neither mean familiarity (Romance = 34.7%, Germanic = 37.0%,
48
49 Japanese = 28.0%) nor median familiarity (Romance = 35.7%, Germanic = 34.3%, Japanese =
50
51 25.0%) is markedly higher for lists in Romance languages than for other languages. We
52
53
54
55
56
57
58
59
60

1
2
3 therefore conclude that more data are needed to understand whether these differences are
4
5 meaningful.
6
7

8 Our final analysis assessed the influence of the familiarity of words in the stimulus
9
10 lists on the preference for lists of familiar word-forms. As a proxy for familiarity, we took the
11
12 median comprehension percentage for individual items as reported in Wordbank. We had
13
14 hypothesized that the familiarity of words chosen in individual studies might differ, and that
15
16 this might explain part of the variability in effect sizes. Indeed, our analyses showed that
17
18 familiarity differed quite substantially between studies: Median familiarity across all word-
19
20 forms in the familiar lists of a given study ranged between a comprehension percentage of
21
22 19% and 69%, and familiarity of the most familiar item ranged between 38% and 100%.
23
24
25
26
27

28 We were subsequently interested in whether these differences would affect effect
29
30 sizes, and whether this would rather be driven by the overall (median) familiarity of the
31
32 word lists, or the familiarity of the most familiar items. Our analyses showed no evidence for
33
34 the latter, *even if we used alternative metrics like items in the highest frequency quantile or*
35
36 *items more than 1 SD more frequent than the mean, and moderate evidence for the former.*
37
38 That is, we found that the overall effect of median familiarity only approached significance,
39
40 and that there was a significant interaction of familiarity with age. The latter was due to a
41
42 higher familiarity effect for younger infants. *Even this effect, however, was not stable in a*
43
44 *leave-one-out analysis, suggesting that it is possibly driven by a few datapoints, or that*
45
46 *overall power is too low. The lack of any effect based on very frequent items thus does not*
47
48 *allow us to support a mechanism in which only very well-known items, the meaning of*
49
50 *which has possibly been encoded, affect infants' familiarity preferences. Instead, our data*
51
52 *show preliminary support for a gradual evidence accumulation process that is reflected in*
53
54
55
56
57
58
59
60

1
2
3 listening times at least in younger infants, who still have limited or less stable word-form
4
5
6 knowledge.

7
8 More research, however, is needed to further examine the impact of the overall
9
10 familiarity of the word list, and the possibility that the benefit of more familiar word-forms
11
12 is age-dependent. First, our chosen proxy is only one of several possible measures of word-
13
14 form familiarity. The definition of familiarity chosen by researchers when building the
15
16 familiar word list could have an impact on infants' recognition and thus on the expected
17
18 effect size. In the studies conducted in our lab which motivated this meta-analysis, we had
19
20 chosen word frequency extracted from infant-directed speech in CHILDES corpora as a
21
22 measure of familiarity. Based on this measure, the three experiments should have given the
23
24 same qualitative results. In particular, the last two experiments (ANONYMIZED FOR REVIEW
25
26 2; Experiments 1 and 2) were carefully constructed, with matched distributions of word
27
28 frequencies, as well as of phone and diphone frequencies. However, based on the CDI
29
30 measure used in this meta-analysis, the two familiar word lists differed in both their
31
32 medians (25% and 19% for experiments 1 and 2, respectively) and their maximum values
33
34 (50% and 38%, respectively), with the median value of the first experiment, which showed
35
36 the familiarity effect, being close to that in Hallé & de Boysson-Bardies (1994) and
37
38 ANONYMIZED FOR REVIEW 1. While the effect of familiarity measured with the CDI was
39
40 inconclusive in the present analysis, it may hint at a problem with the definition of
41
42 familiarity used in our own experiments, which could explain the contradictory results
43
44 obtained in the same laboratory for the same language and age group. Future studies
45
46 should thus assess other familiarity measures, such as word frequency extracted from adult-
47
48 directed speech in CHILDES corpora (as done in ANONYMIZED FOR REVIEW 3) or from newly
49
50 emerging corpora based on day-long recordings of infant input (e.g., Bunce et al., 2020).
51
52
53
54
55
56
57
58
59
60

1
2
3 We also remark that we based our proxy on comprehension data from 12-month-old
4 infants, whereas the preference scores included are based on infants of a wider age range.
5
6 It is possible that results might differ with sufficient data to reflect age more accurately, for
7
8 instance, if we could have calculated, for each infant group tested, our familiarity score
9
10 based on CDIs assessed at that age. Second, considering the rather subtle effect of the
11
12 magnitude of familiarity within familiar word-forms, the number of studies we were able to
13
14 include in our meta-analysis is not large enough to draw conclusions concerning this
15
16 possibility. Finally, the distribution of familiarity scores and age groups is uneven, and both
17
18 of these factors could have decreased the power to detect an effect.
19
20
21
22
23
24

25 Despite these reservations, our tentative results open novel avenues into investigating
26
27 the mechanisms of early word-form acquisition, and even early language acquisition more
28
29 generally. Indeed, one central underlying assumption shared by several accounts of infants'
30
31 early language acquisition is that they accumulate evidence from their linguistic input (e.g.,
32
33 Kuhl et al., 2008; Maye, Werker, & Gerken 2002; Saffran, 2003). In the case of word-forms,
34
35 evidence accumulation takes place over strings that occur frequently in their input. The
36
37 studies included in the present meta-analysis pit "familiar" word-forms which infants should
38
39 have encountered, against "unfamiliar" word-forms which infants should at best have
40
41 encountered rarely. Word familiarity, however, is a graded concept, and our preliminary
42
43 results concerning infants' matching graded sensitivity are in line with the assumption that
44
45 the mechanism of such evidence accumulation should be based on a continuous process
46
47 (see also Tsuji et al., 2017).
48
49
50
51
52
53

54 A final caveat is the potential for publication bias in the present dataset, which was
55
56 manifest in a significant funnel plot asymmetry despite our inclusion of unpublished results.
57
58 Two of these datapoints are based on infants from Romance language backgrounds and
59
60

1
2
3 might thus have contributed to the larger effects found for this group of languages. All
4
5 datapoints are based on infants at age 11 or 12 months, and are thus unlikely to have had a
6
7 large influence on the age effect. Finally, one of the datapoints was based on Welsh, and
8
9 was therefore not included in familiarity analyses. While one of the other two datapoints
10
11 showed a rather high median word familiarity (56%), the other did not (27%), thus overall
12
13 making it unlikely that these data contributed disproportionately to the observed effects. We
14
15 do invite the interested reader, however, to explore the data and bias correction methods
16
17 using the open-access resources provided.
18
19
20
21
22

23 To conclude, our meta-analysis revealed a robust effect of early word-form recognition
24
25 across the languages and stimuli assessed. At the same time, our moderator analyses
26
27 suggest that infant age, native language, as well as stimulus familiarity, might explain some
28
29 of the unexplained variance in the results. We hope that researchers planning future studies
30
31 on word-form recognition can learn from these insights.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Acknowledgements

This research was funded by a doctoral grant from *Ecole des Neurosciences de Paris* to M.J.C., by grants from the *Agence Nationale de la Recherche* (ANR-17-CE28-0007-01 ANR-17-EURE-0017), and by a *The University of Tokyo Excellent Young Researcher* grant to S.T. We would like to thank Marilyn Vihman, Tamar Keren-Portnoy, Tracey Bywater, Daniel Swingley, Pierre Hallé, Camillia Bouchon, Mélanie Hoareau, Thierry Nazzi and Silvana Poltrock for their time, insights, and help in providing us with additional information regarding their experiments, without which this meta-analysis would not have been possible. We are also grateful to the Language Acquisition Department at the Max-Planck Institute for Psycholinguistics and the Baby & Child Research Center Nijmegen for providing us with Dutch CDI data.

Literature

- Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science, 19*(6), 901-917. Doi: 10.1111/desc.12341
- Bunce, J., Soderstrom, M., Bergelson, E., Rosemberg, C., Stein, A., Migdalek, M., & Casillas, M. (2020). A cross-cultural examination of young children's everyday language experiences. Preprint doi: [10.31234/osf.io/723pr](https://doi.org/10.31234/osf.io/723pr)
- Cochrane Collaboration. (2011). *Cochrane Handbook for Systematic Reviews of Interventions* (Version 5.1. 0) 9.6 Investigating heterogeneity <http://handbook-5-1.cochrane.org>.
- Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review, 120*(4), 751–778. Doi: [10.1037/a0034245](https://doi.org/10.1037/a0034245)
- Fenson, L. (2007). *MacArthur-Bates communicative development inventories*. Baltimore, MD: Paul H. Brookes Publishing Company.
- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language 44*(3), 677-694. Doi: [10.1017/S0305000916000209](https://doi.org/10.1017/S0305000916000209)
- Hallé, P. A., & de Boysson-Bardies, B. (1996). The format of representation of recognized words in infants' early receptive lexicon. *Infant Behavior and Development, 19*(4), 463-481. Doi: [10.1016/S0163-6383\(96\)90007-7](https://doi.org/10.1016/S0163-6383(96)90007-7)
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in infancy research*, 5, 69–95.

Jusczyk, P. W. (2002). How infants adapt speech-processing capacities to native-language structure. *Current Directions in Psychological Science*, 11(1), 15-18. Doi: 10.1111/1467-8721.00159

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science*, 277(5334), 1984-1986. Doi: 10.1126/science.277.5334.1984

Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, 18(1), 111–116. Doi: 10.1016/0163-6383(95)90012-8

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 363(1493), 979–1000. Doi: 10.1098/rstb.2007.2154

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), B101–B111. Doi: 10.1016/S0010-0277(01)00157-3

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1), 91-117. Doi: 10.1016/S0010-0277(03)00140-9

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). PRISMA Group: Methods of systematic reviews and meta-analysis: preferred reporting items for systematic reviews

and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology*, 62, 1006-1012.

Ogura, T., Watamaki, T. & Inaba, T. (2016) *The Japanese MacArthur-Bates Communicative Development Inventories*. Kyoto: Nakanishiya co.

Poltock, S., & Nazzi, T. (2015). Consonant/vowel asymmetry in early word-form recognition. *Journal of Experimental Child Psychology*, 131, 135-148. Doi: 10.1016/j.jecp.2014.11.011

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

RStudio Team (2019). *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, 12(4), 110-114. Doi: 10.1111/1467-8721.01243

Swingle, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental science*, 8(5), 432-443. Doi: 10.1111/j.1467-7687.2005.00432.x

Swingle, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536), 3617-3632. Doi: 10.1098/rstb.2009.0107

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, 56(2), 179-191. Doi: 10.1002/dev.21179

Tsuji, S., Fikkert, P., Minagawa, Y., Dupoux, E., Filippin, L., Versteegh, M., ... & Cristia, A. (2017). The more, the better? Behavioral and neural correlates of frequent and infrequent vowel exposure. *Developmental Psychobiology*, 59(5), 603-612. Doi: 10.1002/dev.21534

- 1
2
3 Van Heugten, M., & Johnson, E. K. (2014). Learning to contend with accents in infancy:
4
5 Benefits of brief speaker exposure. *Journal of Experimental Psychology:*
6
7 *General, 143*(1), 340. Doi: 10.1037/a0032192
8
9
10 Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal*
11
12 *of Statistical Software, 36*(3), 1-48. Doi: 10.18637/jss.v036.i03
13
14
15 Vihman, M., & Majorano, M. (2017). The role of geminates in infants' early word
16
17 production and word-form recognition. *Journal of Child Language, 44*(1), 158-184. Doi:
18
19 10.1017/S0305000915000793
20
21
22
23 Vihman, M. M., Nakai, S., DePaolis, R. A., & Hallé, P. (2004). The role of accentual pattern in
24
25 early lexical representation. *Journal of Memory and Language, 50*(3), 336-353. Doi:
26
27 10.1016/j.jml.2003.11.004
28
29
30 Vihman, M. M., Thierry, G., Lum, J., Keren-Portnoy, T., & Martin, P. (2007). Onset of word-
31
32 form recognition in English, Welsh, and English–Welsh bilingual infants. *Applied*
33
34 *Psycholinguistics, 28*(3), 475-493. Doi: 10.1017/S0142716407070269
35
36
37 Von Holzen, K., & Bergmann, C. (2018). A meta-analysis of infants' mispronunciation
38
39 sensitivity development. *Proceedings of the 40th Annual Conference of the Cognitive*
40
41 *Science Society. Austin, TX: Cognitive Science Society, Inc.*
42
43
44
45 Weisleder, A., & Fernald, A. (2013). Talking to children matters early language experience
46
47 strengthens processing and builds vocabulary. *Psychological Science, 24*, 2143–2152.
48
49 Doi: 10.1177/0956797613488145
50
51
52 Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of
53
54 word-object associations by 14-month-old infants. *Developmental Psychology, 34*(6),
55
56 1289–1309. Doi: 10.1037/0012-1649.34.6.1289
57
58
59
60

1
2
3 Wickham. H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version
4
5

6 1.2.1. <https://CRAN.R-project.org/package=tidyverse>
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review Only